Data Mining: Current Applications and Future Possibilities.

Testimony before the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, March 25, 2003

Jen Que Louie, President, Nautilus Systems, Inc.

Thank you Mr. Chairman, Mr. Ranking Member, and other members of the Committee on Government Reform for the opportunity to testify today on the subject of “Data Mining: Current Applications and Future Possibilities.” I will summarize my thoughts briefly in the first pages of my prepared statement and opening remarks, and include more detailed explanation about what data mining is, dispel some of the fallacies about data mining, address what is required for successful data mining analysis and meeting your data mining expectations.

What Is Data Mining?

Depending upon whom you ask, a universal definition of what data mining exactly is can be next to impossible. While the definition seems to be in constant metamorphosis, data mining is an analytic process, whose goal is prediction. The data mining process applies one or more algorithms (computations, queries, links, or sorts) to explore extremely large volumes of data in the hope of discovering patterns and identifying relationships, that were previously unknown, and ultimately make a prediction.

So In a nutshell, data mining is the extraction of knowledge or information from data. Apparent as it may seem at first glance, this concept is a deceptively powerful one. Unlike mere data, knowledge can (1) lead to incisive decisions and (2) reveal previously unknown relationships.

Data Mining Fallacies

The first fallacy is that there are no “data mining tools” simply because, data mining is a process.

With data mining, you just turn loose a plethora of analytic tools, and they will find the answers. While data mining's ability to uncover data patterns can be remarkable, it requires human skills to interpret the results accurately.

The whole data mining process can be autonomous and does not require an analyst once a pattern or set of rules has been identified. This is only true for the specific instance, which the rules were generated from. For example, if a minimum purchase is made with a credit card and that transaction is followed by

the purchase of expensive items, then there is a high probability that the credit card is stolen. This rule is only applicable to identifying possible credit card fraud.

The second fallacy is that the savings realized using data mining pays for itself very rapidly. That depends on what "rapidly" means, along with the cost of the tools, the computational engine, the analyst's time, and your business operation model. Generally speaking, data mining is computationally high and returns a lift of less than one percent to the bottom line.

A third fallacy is that advertised data mining packages are "easy to use and intuitive." Very unlikely, but if you understand the problem you are trying to answer and the tool or tools meet those needs – you are very lucky. Chances are you will require someone with subject matter expertise that is intuitive, analytical, and mathematical to view the overall process, analyze the results, and then make those results actionable events.

Successful Data Mining Analysis

"Data mining is more about letting the data speak for itself," Linoff[1] says.

Data mining differs from other traditional analytical processes by the way data is queried. An analyst using traditional analytical processes usually approaches the problem by constructing a hypothesis or identifying the specific needs to be addressed and using the data available to prove or disprove the hypothesis. Data mining, by comparison, involves targeting a specific problem and using algorithms to form general hypotheses that may expose patterns and relationships that were previously unseen. On the whole, data mining is more predictive in nature than traditional tools that tend to either support or disprove a hypothesis.

An example of how data mining differs from a traditional analysis approach of querying available data. Let us say that a school district administrator's student information database contains historical data about the students enrolled in the district's schools. The administrator wants to know how test scores vary among students from different economic backgrounds. The administrator uses the available data and formulates a query.

A query using a traditional approach may be structured something like: "High school students from low-income households tend to score lower on tests than students from high-income households. Is this true?" The analyst would then generate the appropriate query language for the student information database and generate a report that either supports the hypothesis as correct or wrong. The data might show that students from low-income households do score lower on tests, but overall, the results do not provide much more related information.

Applying the data mining process to the same student information system may identify related information that provides more insight and value. For example, the results might show students from low-income households do tend to score lower on tests; but at the same time, it may also point to other reasons contributing to this pattern. The data mining process might group students who have part-time jobs, come from single-parent households, aren't enrolled in a tutorial program, have a learning disability, have recently moved to the school district, or are frequently absent, as factors that contribute to low test cores. Data mining identified relationships and interdependencies affecting an objective – subject course grades.

Although my example is simple, it illustrates how data mining can unearth unseen and often overlooked pieces of information. This identified information is now actionable information that the administrator can use to apply solutions to the problem.

Putting Data Mining to Use

Over the past few decades we have collected more and more data, to the point that we have no idea what we have. However, with the availability of affordable fast computational processing capabilities in recent years, data mining can make sense of this data for specific business (education, intelligence) purposes.

The one significant shortfall of data mining is that it requires massive quantities of data to be effective. The quality of the prediction is directly proportional to the quality (trustworthiness) and quantity of the data, and the final value of the prediction is dependent on the data mining practitioner's subject matter expertise and insight to deliver actionable results.

When What Is Missing Is What Is Interesting

Sometimes while mining data, the data mining process will kick out an anomaly or flag trends and patterns that should be in the data. The data mining practitioner on his review will either ignore it or follow the thread. Why does it matter?

Case study[2]: The state of California apparently has as many as half a million residents that fail to file state income taxes. Through the use of data mining the California Tax Board is able to identify the fraudulent practice of not filing taxes. This was accomplished by examining past tax returns (historical data) and third-party data (federal W-2 forms). The state of California was able to determine who should have files a return. The use of historical and third parties data makes it relatively easy to determine expected trends and patterns and then detect the absence of them.

The goal here was to identify the absence of lack of a pattern in data, and it is this absence that is flagged as truly interesting. This approach may prove to be especially valuable.

Technology Used By Data Mining Practitioners

The technologies used by data mining practitioners are primarily based on statistical methods such as linear regression, factor analysis, and distribution analysis. The data mining process has extended these foundation algorithms to include more complex and innovative tools that can identify frequencies, associations, temporal events, and patterns from data being mined.

Data mining tools are often categorized by their origins, and usually consist of methods (processes) involving neural networks, clustering, decision trees, classifications, linked lists, correlation, and other numeric methods.

- Neural networks use artificial intelligence (AI) or machine learning processes, which use deductive reasoning, make intelligent estimates, and learn by example.

- Decision trees, originally developed for operations research, provides best-fit logical path solutions.

- Classification and clustering algorithms provide methods for explicit data segmentation. One of the most publicized instances of clustering is in the Geographic Information System (GIS) field, where analysts can, with startling accuracy, identify relative consumption of products within counties and ZIP Codes.

- Linked lists map the relation to any data point; for example, parents to their children, children to their spouses, resulting in another cycle of parents to their children. This is applicable to following money, identify potential laundering, or other types of fraud.

- Correlation matrixes match the same data elements on the X-axis and Y-axis. For example, we list phone numbers on the X and Y-axis, and every intersecting point we enter the number of times (frequency) that the other number calls a particular number. When we graph the matrix, you will be able to visualize the communication relationships. For example, we might see that groups of subscribers by geographic areas call the same pharmacy, transportation schedule recording, or weather report. This information could be utilized to schedule preventative maintenance or upgrades, or evaluate communication equipment utilization by area.

- Data cubes or Multi-Dimensional Database (MDDB), are often categorized as a data mining product. An MDDB is a repository holding aggregations

of data in cells which are the intersection of multiple dimensions (time, geography, product, customer) of the data.

- OLAP (On-Line Analytical Processing) is usually associated with MDDB and has the ability to analyze data across multiple dimensions in a timely manner, in order to support critical decision making.

- Numerical methods in a broad generalization encompass all data mining processes and applications.

- Hybrid Tools
    - CART is a proprietary algorithm developed by Salford Systems, Inc. and uses a nontraditional decision tree methodology. It has a high degree of automation (requires only moderate supervision by the analyst), has the ability to handle arbitrarily complex data structures. Salford Systems claims that a novice-generated first iteration CART model is often as good as a neural net model developed by an expert.

    - Eigen analysis is a multivariable statistical procedure that may be either a prediction or classification technique. It is also capable of discrimination analysis, partial correlation, multiple regression, principal component analysis, and factor analysis.

    - Origami is a hybrid link analysis application developed on the concept of data-cartography, that maps the relationship of data points to each other in a visual representation

    - Information mapping[3] is based on research into how the human mind actually reads processes, remembers, and retrieves information. Nautilus Systems' hyperbolic directory applies information-mapping principles in breaking complex information into its most basic elements and then presents those elements optimally for users. The result is a set of precisely defined information modules that are consistent from author to author and document to document.

    - Intelligent agents are autonomous software computer programs which can dig through data repositories unsupervised and returned with the requested information, monitor for changes in data, or even track who is requesting the data element.

Meeting Your Data Mining Expectations.

Planning is the single most important step in any data mining endeavor. Know and understand what the consumers of your information product need. Then get

the best you can afford hardware and software to enhance the environment that will meet your analysts performance expectations from the outset.

Data mining environments grow more complex and demanding, and sometimes in a short span of time. Design your computational environment with scalability in mind. If your system is not easily scalable, you will have serious performance bottlenecks and major upgrading costs later.

Understand your consumer's operational and information needs. The success of your data mining efforts depends on how well you respond to the dynamics of your consumer's environment.

Time is your worst enemy and faster computers do not necessarily translate into faster insight. Remember, it take a woman nine months to produce a baby, and no matter how hard you try you can not get nine women to make a baby in one month. Allow time for quality assurance and review before delivery of the information product.

Don't underestimate the need for training. Even the brightest science and international law graduates can be shockingly unprepared to take advantage of the tools you are providing them. Don't assume a level of expertise they may not have. Be prepared to provide a substantial amount of training, especially in the area of turning strategic questions into structured queries.

Summation

Data analysis is concerned with the discovery and examination of patterns and associations found in the data. There are various ways to achieve this objective, but all share the fundamental notion that patterns to be examined are present in the data. Also remember that what is not in the data can be just as interesting and in certain situations more useful to know.

Data mining is a process that involves multiple analytic tools and methodologies, driven by the needs of that information product's consumer.

The quality of the information product is directly proportional to the trustworthiness and quantity of the data available.

The confidence of the prediction is dependent on the data mining practitioner's subject matter expertise and insight to deliver actionable results.

The data mining process is highly computational and takes time. Therefore planning the approach and the selection of tools is influenced by the needs of the consumer.

[1] Michael Berry and Gordon Linoff are co-authors of several books on data mining, including “Mastering Data Mining” and “Data Mining Techniques For Marketing, Sales, and Customer Support.” They are also the founders of Data Miners (http://www.data-miners.com), a consultant agency specializing in data mining training and planning.

[2] Round table discussion at, Salashan ’99 High Performance Computing Conference, statement made by Dr. Inderpal Bhandari, founder and CEO of Virtual Gold, Inc., and internationally recognized expert in data mining.

[3] Robert E. Horn, while a student at Harvard and Columbia Universities, conducted research about how readers deal with large amounts of information. This resulted in a standard approach for communicating information, which is based on learning theory, human factors engineering, and cognitive science.